provides a bias towards a bimodal distribution of the distances between the i-th and i + 4th side chain units. The definition of these potentials mimics some of the most general structural regularities seen in all folded proteins. They also provide a bias against nonphysical local conformations in the unfolded state.

Similar to the short-range interactions, there are sequence-specific and generic terms of the model tertiary interaction scheme. The pairwise contact potentials and the model of hydrophobic burial potentials of mean force derived from the statistics of the structural database do not require additional discussion. The procedures of derivation and implementation of such potentials are rather standard and commonly used in all reduced models of proteins.[15-17,21,25,26] Such statistical potentials encode some interaction preferences in real proteins. In the majority of cases, they are accurate enough to select a proper fold for a given sequence from a collection of other folds of natural proteins. However, here, the requirements are more stringent. The proper fold must be selected from a much larger number of conformations, most of them never observed in real proteins (but possible in the model due to the reduced representation). Thus, it is important to construct a generic potential that provides a bias toward protein-like tertiary interaction patterns. Such patterns could be postulated as a generalization of structural regularities seen in known protein structures. An important feature of all protein structures is the very regular network of main chain hydrogen bonds. Our model lacks an explicit protein backbone. Nevertheless, an analysis of protein structures shows that the presence of hydrogen bonds between residues translates with high reproducibility into a pattern of contacting side chains. Indeed, the string of hydrogen bonds along a helix implies the existence of continuous or almost continuous strings of side group contacts along the helix surface. Similarly, a string of hydrogen bonds in a β-hairpin implies two strings of side group contacts, one on each side of the hairpin. Thus, a bias towards such a string (*see* equation 5, above, and the associated discussion of the potential) could be used as an ersatz copy of the hydrogen bond interactions. Furthermore, such strings of contacts lead to a

characteristic pattern of side chain contacts. The generic potential given in equation 6, above, provides a bias towards the most general feature of such patterns.[16,18]

Angular packing preferences for various types (hydrophobic or hydrophilic) of residues also could be used as a bias toward protein-like side chain packing patterns (*see* equation 7, above, and the associated description of this term).

Such a defined model of the force field can be tested and the relative weights of the sequence-specific versus generic terms adjusted by a trial and error method. Here, a long series of isothermal simulations of various proteins was performed. While the native-like structures were sometimes obtained only for very simple, small proteins, the accuracy (measured as cRMSD from native) of the emerging elements of secondary and super-secondary structure elements (helices, helical hairpins, β-hairpins or α-β-α motifs) could be used as a convenient criterion.

This force field alone, however, is not accurate enough for reproducible folding simulations of the majority of (even small) proteins. At the same time, it discriminates against a vast majority of nonsensical conformations, and the native-like structures always belong to a relatively small number of low energy conformations. Thus, when some long-range constraints of experimental origin are superimposed on top of this force field, the native-like conformation can easily be obtained in Monte Carlo simulations, as described below.

Implementation of the Constraints

Encoding short-range conformational propensities

In testing structure assembly using the methods described herein, knowledge of secondary structure[37] is used in the form of the following three-letter code: E-extended; H-helix; and (-) everything else. This three-letter code is then translated onto a set of biases towards a corresponding range of local intrachain distances and angular correlations. Only E and H states have some conformational biases, and their definitions are geometrically very permissive. The set of secondary structural constraints are as follows:

1. An H-state cannot be hydrogen bonded to an E-state. When detected, such bonds are ignored and do not contribute to the conformational energy.

2. A residue in a continuous stretch of H-states can hydrogen bond only to residues i - 3 and i + 3. Note that hydrogen bonds associated with Cα's or side chains represent the canonical helix pattern.

3. The system gains an additional energy equal to $-\varepsilon_{gen}$ (over the previously defined generic contributions, and ε_{gen} is of the same exact value as that used in the definition of various generic germs of the model force field in all the following cases (a-c): As shown in Figure 6 for helical type states when:

$$\text{for} \quad r^2_{i,i+4} < 33 \qquad 13(a)$$

a) residues i + 1 and i + 2 are assigned as helical if $(\mathbf{v}_i \cdot \mathbf{v}_{i+2}) < 0$

b) residues i + 2 and i + 3 are assigned as helical if $(\mathbf{v}_{i+1} \cdot \mathbf{v}_{i+3}) < 0$

c) residues i + 1, i + 2 nd i + 3 are assigned as helical if $(\mathbf{v}_i \cdot \mathbf{v}_{i+3}) < 0$.

As shown in Figure 7, for expanded states when

$$\text{for } 48 < r^2_{i,i-4} < 145 \qquad 13(b)$$

a) residues i + 1 and i + 2 assigned as extended if $(\mathbf{v}_i \cdot \mathbf{v}_{i+2}) < 8$

b) residues i + 2 and i + 3 assigned as extended if $(\mathbf{v}_{i+1} \cdot \mathbf{v}_{i+3}) < 8$

c) residues i + 1, i + 2 and i + 3 assigned as extended if $(\mathbf{v}_i \cdot \mathbf{v}_{i+3}) < 0$.

The set of conditions given in equations 13a and 13b, above, describe various geometrical boundaries for the local conformation of the model chain that are characteristic for helical and expanded states, respectively. In each case, they were split into three sets of conditions to make the energy landscape as smooth as possible (otherwise, a single condition could be applied). In the present realization, the model system gains some energetic stabilization when even a nucleus of a helix or extended state forms. On the other hand, the conditions are rather permissive, allowing substantial fluctuations of the secondary structure without an energetical